

Large Deviations for Random Trees

Yuri Bakhtin¹, Christine Heitsch²

February 2, 2008

Abstract

We consider large random trees under Gibbs distributions and prove a Large Deviation Principle (LDP) for the distribution of degrees of vertices of the tree. The LDP rate function is given explicitly. An immediate consequence is a Law of Large Numbers for the distribution of vertex degrees in a large random tree. Our motivation for this study comes from the analysis of RNA secondary structures.

Keywords: *random trees, Gibbs distributions, large deviations, RNA secondary structure*

1 Introduction

In this note, we prove a Large Deviation Principle (LDP) for two models of equilibrium statistical mechanics. In both cases, we consider a set of trees on N vertices and we define the Gibbs distribution associated to a certain energy function on that set. The main goal of our work is to study some typical features of large random trees ($N \rightarrow \infty$) under these distributions.

Here, we provide rigorous proofs for the LDP results announced in [BH]. As discussed there, our results are motivated by, and have applications to, the branching of RNA secondary structures. The trees we consider are a useful abstraction of these biological structures (see [Heib, Heia] for references on this connection) as well as relatively straightforward to analyze mathematically. In this simplified model of RNA folding, we can address the interplay between entropy and energy in determining a “typical” branching configuration. We find that, due

¹School of Mathematics, Georgia Tech, Atlanta GA, 30332-0160;
email: bakhtin@math.gatech.edu,

²School of Mathematics, Georgia Tech, Atlanta GA, 30332-0160;
email: heitsch@math.gatech.edu

to the entropy factor, the typical configurations in our model differ from the arrangements which have minimal energy in interesting ways.

Our mathematical results support and extend recent developments in RNA secondary structure prediction (reviewed in [Mat06, MT06]) which broaden the focus beyond simply finding a structure with minimal free energy. In particular, we prove a Law of Large Numbers for the degree frequencies in our large random trees, and find that the most common trees are not the minimizers of the associated energies. This highlights the limitations of prediction methods focused solely on energy minimization and the significance of entropy considerations in computational structural biology.

2 Models and results

In this section we describe our models and state the results. The proofs are given in the next section.

2.1 Labeled trees

In our first model we fix a natural number $D \geq 2$ and for each $N \in \mathbb{N}$ consider the set $\mathbb{T}_N(D)$ of labeled trees on $N \in \mathbb{N}$ vertices such that the degree of each vertex does not exceed D . To define Gibbs distributions on $\mathbb{T}_N(D)$ we need a function $c : \{1, \dots, D\} \rightarrow \mathbb{R}$ which plays the role of the energy associated with the degree of a vertex.

To each of the trees T in $\mathbb{T}_N(D)$ we associate the energy

$$H(T) = \sum_{j=1}^N c(d_j(T)) = \sum_{k=1}^D c(k) \chi_k(T), \quad (1)$$

where $d_j(T)$ denotes the degree of the j -th vertex, and $\chi_k(T)$ is the number of vertices of degree k in T . Now the Gibbs probability measure on $\mathbb{T}_N(D)$ associated with H is given by

$$P_N\{T\} = \frac{e^{-\beta H(T)}}{Z_N}, \quad T \in \mathbb{T}_N(D),$$

where $\beta > 0$ is the inverse temperature parameter and

$$Z_N = \sum_{T \in \mathbb{T}_N} e^{-\beta H(T)} \quad (2)$$

is the partition function.

Our first result is an LDP for the degree distribution of random labeled trees under measures P_N introduced above.

Let us recall that a sequence of probability measures $(\mu_N)_{N \in \mathbb{N}}$ on a compact metric space (E, ρ) satisfies an LDP with a lower-semicontinuous nonnegative rate function $I : E \rightarrow \mathbb{R}$ if

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \ln \mu_N(C) \leq -I(C), \quad \text{for any closed set } C \subset E,$$

and

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \ln \mu_N(O) \geq -I(O), \quad \text{for any open set } O \subset E,$$

where for $U \subset E$,

$$I(U) = \inf_{p \in U} I(p).$$

See [Ell06, Section II.3] or [DZ98, Section 1.2] for further details.

Informally, an LDP means that if we consider random variables X_N with distribution μ_N , then for all p and large N we have

$$\mu_N\{X_N \approx p\} \approx e^{-NI(p)}.$$

In particular, if the minimal value 0 is attained by I at a unique point p^* then for any neighborhood O of p^* , $\mu_N(O^c)$ decays exponentially in N . This can be restated as a Law of Large Numbers with exponential convergence in probability to the limit point p^* .

We can view (χ_1, \dots, χ_D) as a random vector defined on the probability space $\mathbb{T}_N(D)$ equipped with the Gibbs measure P_N . We would like to study the frequencies of vertex degrees, so for each N we introduce a probability measure ν_N on $[0, 1]^D$ defined as the distribution of the random vector $\frac{1}{N}(\chi_1, \dots, \chi_D)$ under P_N . It is natural to formulate an LDP for ν_N on the set

$$\mathcal{M} = \left\{ p \in [0, 1]^D : \sum_{k=1}^D p_k = 1, \sum_{k=1}^D k p_k = 2 \right\}$$

equipped with Euclidean distance. (Notice that \mathcal{M} is nonempty if $D \geq 2$.) Though the random vector $\frac{1}{N}(\chi_1, \dots, \chi_D)$ does not belong to \mathcal{M} , it is asymptotically close to \mathcal{M} :

$$\sum_{k=1}^D \frac{\chi_k}{N} = 1, \quad \sum_{k=1}^D k \frac{\chi_k}{N} = 2 - \frac{2}{N}.$$

So instead of formulating an LDP for the sequence of random vectors $\frac{1}{N}(\chi_1, \dots, \chi_D)$, we shall formulate and prove an LDP for a sequence of random vectors that is close to it and belongs to \mathcal{M} .

To define the rate function, we introduce $J : \mathcal{M} \rightarrow \mathbb{R}$ via

$$J(p) = -h(p) + \beta E(p) + G(p),$$

where

$$h(p) = - \sum_{k=1}^D p_k \ln p_k$$

is the entropy of the probability vector $p = (p_1, \dots, p_D)$,

$$E(p) = \sum_{k=1}^D p_k c(k)$$

is the energy associated with p , and $G(p)$ is defined by

$$G(p) = \sum_{k=1}^D p_k \ln((k-1)!). \quad (3)$$

In Section 3, we shall see that the function G appears naturally in the analysis of random trees.

The function J is strictly convex down and continuous on \mathcal{M} . Therefore, it attains its minimal value at a uniquely defined point $p^* \in \mathcal{M}$. Consider now

$$I(p) = J(p) - J(p^*). \quad (4)$$

It is easy to see that I is bounded, convex and continuous on \mathcal{M} .

For a measure Q on $[0, 1]^D \times \mathcal{M}$ we define $Q^{(1)}$ and $Q^{(2)}$ as the marginal distributions of Q on $[0, 1]^D$ and \mathcal{M} respectively.

Theorem 1 *There is a sequence of probability measures $(Q_N)_{N \in \mathbb{N}}$ defined on $[0, 1]^D \times \mathcal{M}$ with the following properties.*

1. *For each N , we have $Q_N^{(1)} = \nu_N$.*
2. *For each N ,*

$$Q_N \left\{ (x, y) \in [0, 1]^D \times \mathcal{M} : \sum_{k=1}^D |x_k - y_k| > \frac{2}{N} \right\} = 0.$$

3. *The sequence $(Q_N^{(2)})_{N \in \mathbb{N}}$ satisfies an LDP on \mathcal{M} with the rate function I defined in (4).*

Remark 1 This theorem says that although the random vector χ/N does not belong to \mathcal{M} , one can find another random vector that is, on the one hand, very close to χ/N and on the other hand belongs to \mathcal{M} and satisfies the LDP.

Theorem 1 immediately implies the following Law of Large Numbers:

Corollary 1 As $N \rightarrow \infty$,

$$\left(\frac{\chi_1}{N}, \dots, \frac{\chi_D}{N}\right) \rightarrow p^*$$

in probability.

Remark 2 The statements above show that with high probability the degree frequencies are close to p^* . Note that in most cases the minimum of the energy E on \mathcal{M} is not attained at p^* .

2.2 Plane trees

We now consider a similar model for plane trees (sometimes also called ordered trees). These are rooted trees such that subtrees at any vertex are linearly ordered, see e.g. [Sta99]. We redefine the notation introduced in the previous section. We fix a number $D \in \mathbb{N}$ and for each $N \in \mathbb{N}$ let $\mathbb{T}_N(D)$ denote the set of ordered trees on $N \in \mathbb{N}$ vertices such that the branching (i.e. the number of children) at each vertex does not exceed D . The energy of each vertex depends only on its branching and is given by a function $c : \{0, 1, \dots, D\} \rightarrow \mathbb{R}$. With each tree $T \in \mathbb{T}_N(D)$ we associate the energy

$$H(T) = \sum_{k=0}^D c(k) \chi_k(T), \quad (5)$$

where $\chi_k(T)$ is now the number of vertices with k children in T . The Gibbs probability measure on $\mathbb{T}_N(D)$ associated with H is given by

$$P_N\{T\} = \frac{e^{-\beta H(T)}}{Z_N}, \quad T \in \mathbb{T}_N(D),$$

where $\beta > 0$ is the inverse temperature and Z_N is a normalizing constant.

For each N , we introduce a probability measure ν_N on $[0, 1]^{D+1}$ defined as the distribution of the random vector $\frac{1}{N}(\chi_0, \chi_1, \dots, \chi_D)$ under P_N .

We redefine \mathcal{M} to be

$$\mathcal{M} = \left\{ p \in [0, 1]^{D+1} : \sum_{k=0}^D p_k = 1, \sum_{k=0}^D k p_k = 1 \right\}.$$

To formulate an LDP for this model we define $J : \mathcal{M} \rightarrow \mathbb{R}$ via

$$J(p) = -h(p) + \beta E(p),$$

where

$$h(p) = - \sum_{k=0}^D p_k \ln p_k$$

is the entropy of the probability vector $p = (p_0, p_1, \dots, p_D)$, and

$$E(p) = \sum_{k=0}^D p_k c(k)$$

is the energy associated with $p \in \mathcal{M}$.

As in the first model, the function J attains its minimum on \mathcal{M} at a unique point that we denote by p^* . Let

$$I(p) = J(p) - J(p^*). \quad (6)$$

This function will play the role of the rate function. Notice that in the case of plane trees it does not involve the function $G(p)$ that appeared in the construction of the rate function for the case of labeled trees.

For a measure Q on $[0, 1]^{D+1} \times \mathcal{M}$ we define $Q^{(1)}$ and $Q^{(2)}$ as the marginal distributions of Q on $[0, 1]^{D+1}$ and \mathcal{M} respectively.

Theorem 2 *There is a sequence of probability measures $(Q_N)_{N \in \mathbb{N}}$ defined on $[0, 1]^{D+1} \times \mathcal{M}$ with the following properties.*

1. *For each N , we have $Q_N^{(1)} = \nu_N$.*
2. *For each N ,*

$$Q_N \left\{ (x, y) \in [0, 1]^{D+1} \times \mathcal{M} : \sum_{k=0}^D |x_k - y_k| > \frac{1}{N} \right\} = 0.$$

3. *The sequence $(Q_N^{(2)})_{N \in \mathbb{N}}$ satisfies an LDP on \mathcal{M} with the rate function I defined in (6).*

An immediate consequence is the following Law of Large Numbers:

Corollary 2 *As $N \rightarrow \infty$,*

$$\left(\frac{\chi_0}{N}, \frac{\chi_1}{N}, \dots, \frac{\chi_D}{N} \right) \rightarrow p^*$$

in probability.

3 Proofs

We start with the proof of Theorem 1, adopting the notation and setting for labeled trees from Section 2.1.

The crucial fact for our analysis is the following formula for the number of trees on N vertices with degrees d_1, \dots, d_N :

$$\binom{N-2}{d_1-1, d_2-1, \dots, d_N-1}$$

if $d_1 + \dots + d_N = 2N - 2$, and 0 otherwise, see [Moo70, Formula (2.1)]. Therefore, the total number of N -trees T with $\chi(T) = (n_1, \dots, n_D)$ is given by

$$\binom{N-2}{\underbrace{0, \dots, 0}_{n_1}, \underbrace{1, \dots, 1}_{n_2}, \dots, \underbrace{D-1, \dots, D-1}_{n_D}} \binom{N}{n_1, \dots, n_D} = \frac{(N-2)!}{(2!)^{n_3} \dots ((D-1)!)^{n_D}} C(N, n),$$

where $C(N, n) = \binom{N}{n_1, \dots, n_D}$. All these trees T have the same energy $H(T)$, so that

$$P_N \left\{ \frac{\chi(T)}{N} = \frac{n}{N} \right\} = \frac{e^{-NF(\frac{n}{N})} C(N, n)}{Z_N}, \quad (7)$$

where Z_N is defined in (2), and we notice that

$$Z_N = \sum_{\substack{n_1 + \dots + n_D = N \\ n_1 + \dots + Dn_D = 2N-2}} e^{-NF(\frac{n}{N})} C(N, n),$$

and

$$F(p) = \beta E(p) + G(p) = \beta \sum_{k=1}^D c(k) p_k + \sum_{k=1}^D \ln((k-1)!) p_k, \quad p \in [0, 1]^D,$$

with $G(p)$ defined in (3).

Our plan is to use the LDP for multinomial distribution that manifests itself in coefficients $C(N, n)$ in the r.h.s. of (7), and then apply a version of Varadhan's lemma for Gibbs transformation via the exponential factor $e^{-NF(\frac{n}{N})}$.

We start with the family of distributions μ_N on \mathcal{M} defined by

$$\mu_N \left\{ \left(\frac{n_1}{N}, \dots, \frac{n_D}{N} \right) \right\} = \begin{cases} \frac{C(N, n)}{Z'_N}, & \text{if } \left(\frac{n_1}{N}, \dots, \frac{n_D}{N} \right) \in \mathcal{M} \\ 0, & \text{otherwise} \end{cases},$$

where

$$Z'_N = \sum_{n/N \in \mathcal{M}} C(N, n).$$

Lemma 1 *The sequence of measures $(\mu_N)_{N \in \mathbb{N}}$ satisfies an LDP on \mathcal{M} with rate function I_1 defined by*

$$I_1(p) = h^* - h(p),$$

where

$$h^* = \sup_{p \in \mathcal{M}} h(p).$$

Proof. The proof of this lemma literally repeats that of Sanov's theorem (an LDP for the multinomial distribution, see [DZ98, Theorem 2.1.10]). It is based on the formula:

$$\frac{1}{N} \ln C(N, n) = - \sum_{k=1}^D \frac{n_k}{N} \ln \frac{n_k}{N} + O\left(\frac{\ln N}{N}\right), \text{ as } N \rightarrow \infty,$$

which holds true uniformly in n , see e.g. [Ell06, Lemma I.4.4].

Let us now introduce the Gibbsian weight

$$q_N\left(\frac{n}{N}\right) = e^{-NF\left(\frac{n}{N}\right)},$$

and a new family of measures λ_N on \mathcal{M} :

$$\lambda_N\left\{\frac{n}{N}\right\} = \frac{q_N\left(\frac{n}{N}\right) \mu_N\left\{\frac{n}{N}\right\}}{Z_N''}, \quad \text{for } \frac{n}{N} \in \mathcal{M},$$

where

$$Z_N'' = \sum_{\frac{n}{N} \in \mathcal{M}} q_N\left(\frac{n}{N}\right) \mu_N\left\{\frac{n}{N}\right\} = \int_{\mathcal{M}} e^{-NF(p)} \mu_N(dp).$$

In other words,

$$\lambda_N(dp) = \frac{e^{-NF(p)} \mu_N(dp)}{\int_{\mathcal{M}} e^{-NF(p)} \mu_N(dp)}.$$

Let us also denote $J_1(p) = F(p) + I_1(p)$ and $J_{1,*} = \inf_{p \in \mathcal{M}} J_1(p)$.

Lemma 2 *The sequence of measures $(\lambda_N)_{N \in \mathbb{N}}$ satisfies an LDP on \mathcal{M} with rate function I_2 given by $I_2(p) = J_1(p) - J_{1,*}$.*

Proof. This lemma follows directly from a variant of Varadhan's lemma for Gibbs transformations (Theorem II.7.2 in [Ellis]).

Remark 3 Notice that $I_2(p) = I(p)$ for all $p \in \mathcal{M}$. So we have proven the desired LDP on \mathcal{M} for $(\lambda_N)_{N \in \mathbb{N}}$, and in order to prove Theorem 1 we shall have to compare λ_N to ν_N .

Proof of Theorem 1. We consider the distribution P_N on $\mathbb{T}_N(D)$, so that $\frac{x}{N}$ is distributed according to ν_N . For each x that belongs to the support of ν_N we introduce the set

$$R(x) = \left\{ y \in \mathcal{M} : y_k = \frac{m_k}{N}, m_k \in \mathbb{Z}, k = 1, \dots, D, \right. \\ \left. \text{and } \sum_{k=1}^D |x_k - y_k| = \frac{2}{N} \right\}.$$

It is easy to see that $1 \leq |R(x)| \leq D^2$ for all x , where $|R|$ denotes the number of elements in R .

Let us now define the measure Q_N . We start with random variables χ/N , and define a random vector Y so that, given χ/N , the conditional distribution of Y is uniform on $R(\chi/N)$. Now Q_N denotes the joint distribution of χ/N and Y . Clearly, the first two desired properties of Q hold true by the definition of Q_N . The third one follows from Lemma 2 and the following statement claiming that measures $Q_N^{(2)}$ and λ_N differ by a subexponential factor, thus obeying an LDP with the same rate function:

Lemma 3 *There is a constant $C > 0$ such that for all N and all sets $U \subset \mathcal{M}$,*

$$\frac{1}{CN^4} \leq \frac{Q_N^{(2)}(U)}{\lambda_N(U)} \leq CN^4.$$

This lemma is a straightforward consequence of the following fact: there is a constant K such that if $|n_1 - n'_1| + \dots + |n_D - n'_D| = 2$ then

$$\frac{1}{KN^2} \leq \frac{e^{-NF(\frac{n}{N})}C(N, n)}{e^{-NF(\frac{n'}{N})}C(N, n')} \leq KN^2.$$

The proof of Theorem 2 is essentially the same. It is based on the following expression for the number of ordered trees of order N with n_k nodes having k children:

$$\frac{1}{N} \binom{N}{n_0, n_1, n_2, \dots}$$

if $n_1 + 2n_2 + \dots = N - 1$, and 0 otherwise (see e.g. Theorem 5.3.10 in [Sta99]).

References

- [BH] Yuri Bakhtin and Christine Heitsch. Large deviations for random trees and the branching of RNA secondary structures. Submitted to Bulletin of Mathematical Biology.
- [DZ98] Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 1998.
- [Ell06] Richard S. Ellis. *Entropy, large deviations, and statistical mechanics*. Classics in Mathematics. Springer-Verlag, Berlin, 2006. Reprint of the 1985 original.

- [Heia] Christine E. Heitsch. Combinatorial insights into RNA secondary structures. In preparation for the Journal of Computational Biology.
- [Heib] Christine E. Heitsch. Combinatorics on plane trees, motivated by RNA secondary structure configurations. Submitted to SIAM Journal on Discrete Mathematics.
- [Mat06] David H. Mathews. Revolutions in RNA secondary structure prediction. *J Mol Biol*, 359(3):526–32, Jun 9 2006.
- [Moo70] J. W. Moon. *Counting labelled trees*, volume 1969 of *From lectures delivered to the Twelfth Biennial Seminar of the Canadian Mathematical Congress (Vancouver)*. Canadian Mathematical Congress, Montreal, Que., 1970.
- [MT06] David H. Mathews and Douglas H. Turner. Prediction of RNA secondary structure by free energy minimization. *Curr Opin Struct Biol*, 16(3):270–278, 2006.
- [Sta99] Richard P. Stanley. *Enumerative combinatorics. Vol. 2*, volume 62 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1999. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin.